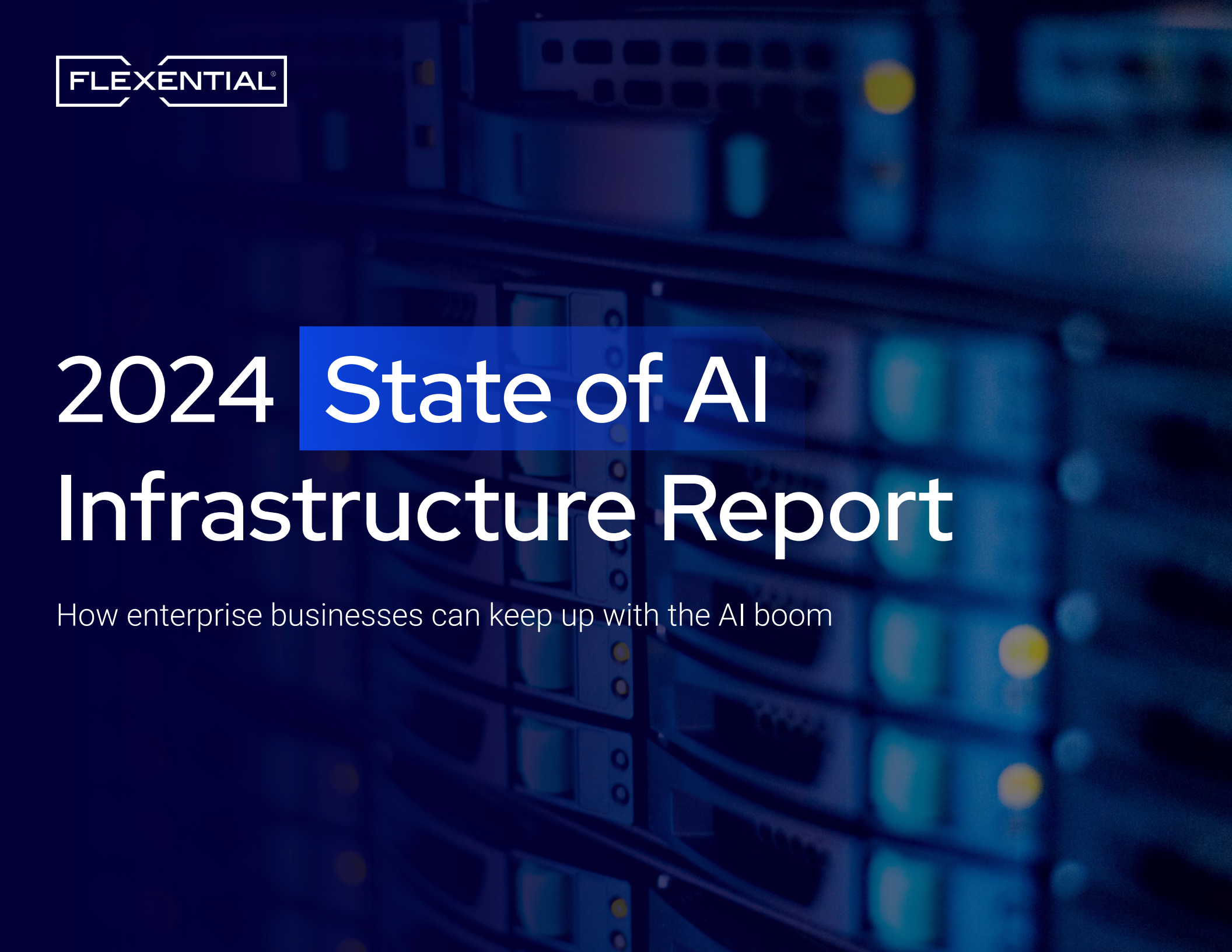


FLEXENTIAL®

# 2024 State of AI Infrastructure Report

How enterprise businesses can keep up with the AI boom

# Table of contents



# Organizations across industries have announced ambitious AI roadmaps. But is their existing IT infrastructure up to the challenge?

Few cloud deployments or on-premises data centers were designed with high-density AI workloads or latency-sensitive AI applications in mind. As the generative AI (gen AI) boom continues to drive heavy investments in this compute-intensive technology, IT leaders need to **quickly adapt and expand current infrastructure** to meet ever-evolving AI needs.

To better understand how organizations are responding to this and other pressures, Flexential surveyed 350 IT leaders at organizations with over $100 million in annual revenue, including 100 respondents at organizations with over $2 billion in annual revenue.

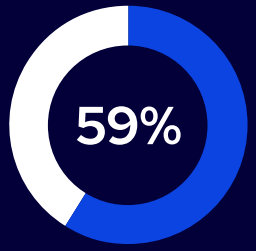
To overcome these challenges, organizations are experimenting with new solutions — for example, reducing latency through strategic deployment of AI workloads to **third-party colocation data centers**, which can process data closer to the network's edge or closer to their existing data sources.

But experimentation alone is not enough. To harness the AI revolution's full potential, organizations must level up their IT infrastructure to match their AI ambitions and vision. This includes exploring and embracing the complexity of AI infrastructure challenges by taking a **more strategic, proactive approach to AI workload deployment** informed by third-party expertise.

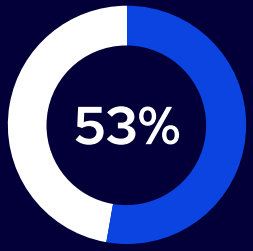
**We found that while IT leaders are enthusiastic and cautiously optimistic about their organizations' AI plans, they also face significant roadblocks, including:**

- Challenges networking and scaling data centers
- Skills gaps related to specialized infrastructure
- Increased data privacy and security risks
- Growing concerns about IT infrastructure sustainability
- C-suite disconnection from challenges on the ground

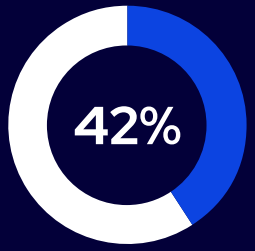
# Key findings



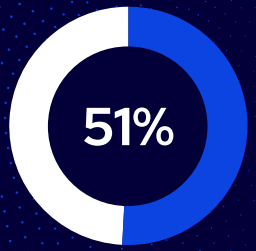

of respondents from organizations with AI roadmaps said **increasing IT infrastructure investments** was an element of that roadmap



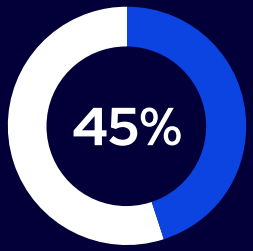

of respondents reported **skills gaps or staffing shortages** related to the management of specialized computing infrastructure



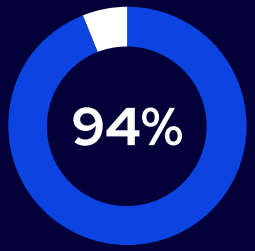

of respondents whose organizations have pulled an AI workload back from public cloud said it was due to **data privacy and security concerns**




of respondents are addressing performance issues by using **third-party colocation data centers** to process data closer to the edge of the network




of respondents from organizations with AI roadmaps said failing to meet the goals in their roadmap would **affect their ability to innovate**




of respondents would pay more for data centers or third-party cloud vendors to **use clean or renewable energy** and/or buy credits to **offset their carbon footprints**

# IT leaders feel excited about their organizations' ambitious AI roadmaps, but ability to execute remains in question

## Data center infrastructure is a top priority in organizations' AI roadmaps

While the current AI boom is mostly driven by gen AI solutions, machine learning applications have been available for years, and are currently deployed across many industries, including finance and healthcare. Nearly all respondents (99%) said their organization has a **documented AI roadmap** — likely reflective of these earlier use cases.

Fifty-nine percent of respondents whose organizations had AI roadmaps said **increasing infrastructure investments** was a part of that roadmap — the most popular answer chosen [Fig. 1]. Investing in stronger cybersecurity protections for AI applications came in second at 54%.

FIG. 1

**Which of the following are elements of your organization's AI roadmap?** Select all that apply.

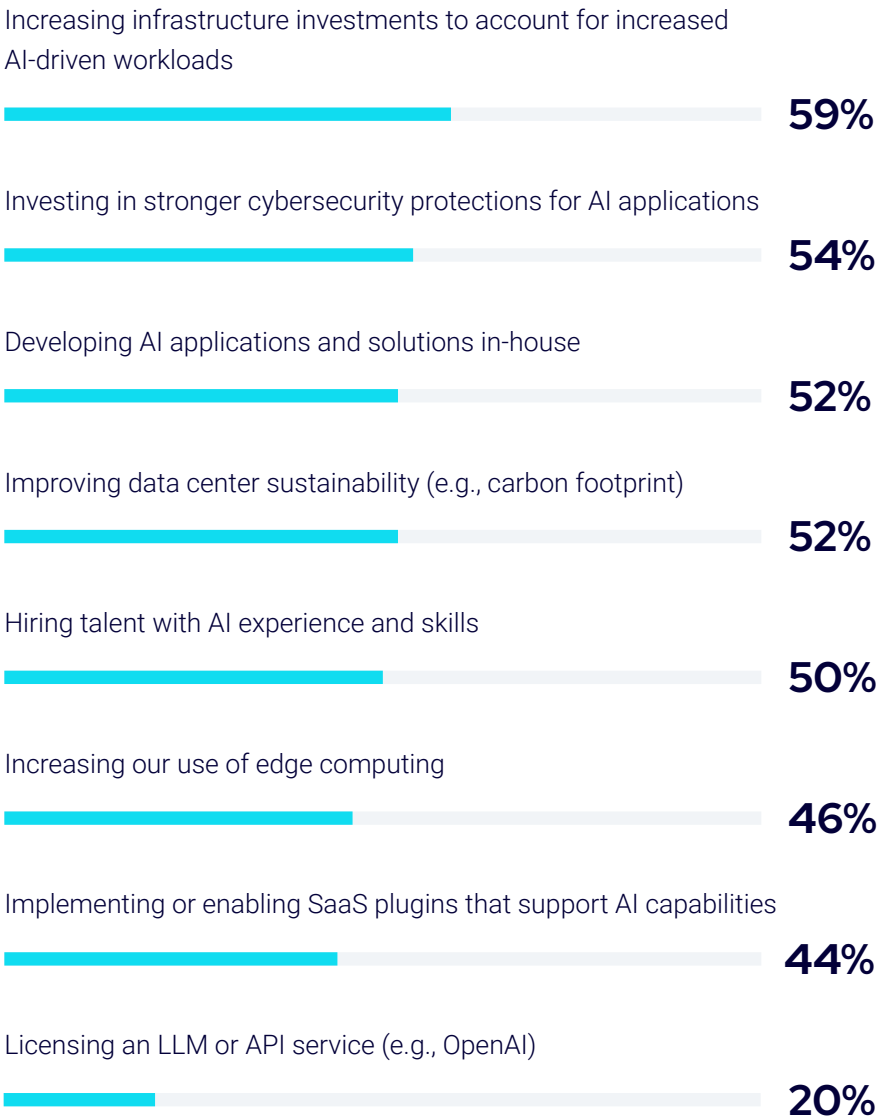


FIG. 2

## Which of the following groups or individuals are the driving force behind your organization's decision to adopt AI-driven applications?

Select no more than three.

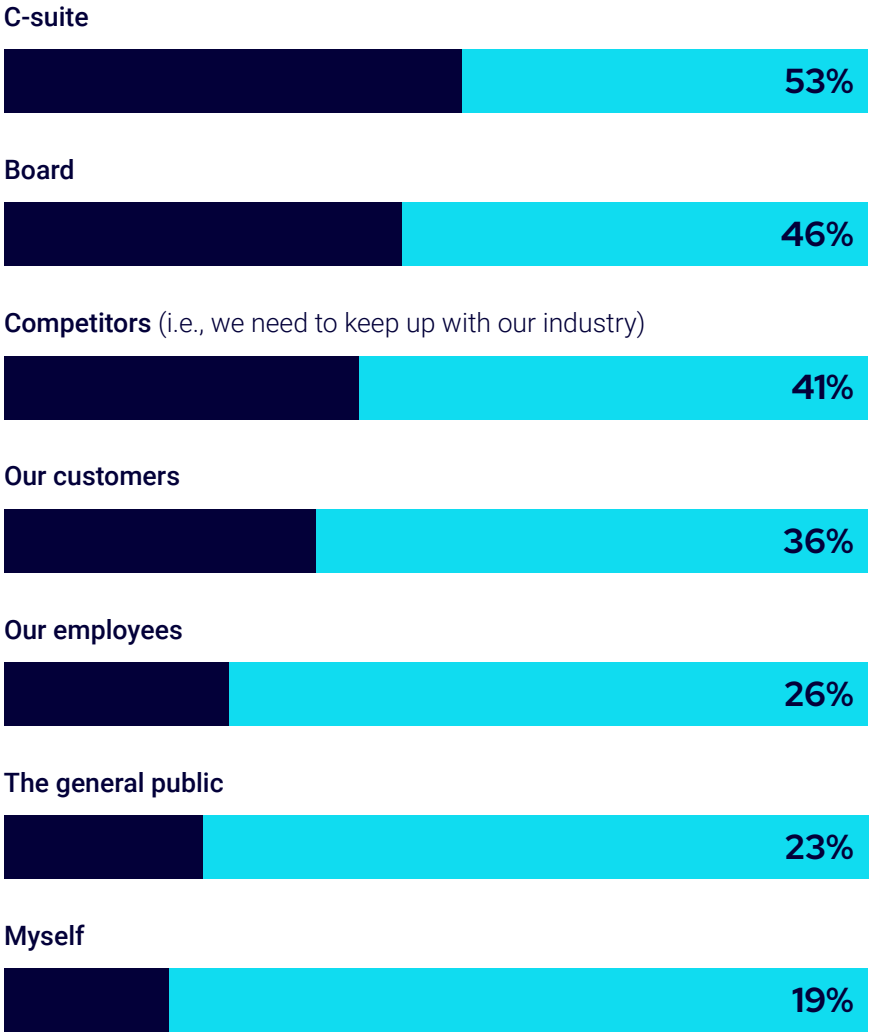



FIG. 2A

## Which of the following groups or individuals are the driving force behind your organization's decision to adopt AI-driven applications?

Select no more than three.

| Company revenue | Percentage of respondents selecting "C-suite" |
|---|---|
| $101M - $500M | 51% |
| $501M - $2B | 50% |
| > $2B | 59% |

## AI is a board-level conversation, and IT leaders are under increased scrutiny

AI investments are a **top-down initiative** at most organizations. Over half of respondents (53%) said the C-suite was one of the top three driving forces behind AI adoption, and almost half (46%) identified the board as a driving force [Fig. 2].

This trend was more pronounced for larger organizations. Fifty-nine percent of respondents at organizations with more than $2 billion in annual revenue said the **C-suite was a driving force behind AI adoption**, compared to roughly 50% of respondents at smaller organizations [Fig. 2a].

C-suite and board attention could prove a **double-edged sword**: It means more support (and likely more resources) for AI initiatives, but more scrutiny on AI-related investments as well.

**Nearly all respondents (93%)** said they agreed or somewhat agreed with the statement:

## “Compared to five years ago, there’s a greater expectation that IT leaders in my organization minimize time-to-revenue for AI-driven IT infrastructure.”

### IT leaders are enthusiastic and bought in on AI plans, but less confident about execution

Respondents were generally enthusiastic about their organizations’ AI efforts. Nearly three-quarters (73%) say they’re **excited** about AI initiatives in their organization, and almost half (49%) say they feel **inspired**. Only small minorities of IT leaders cite negative feelings like nervousness (16%) or a sense of overwhelm (12%) [Fig. 3].

However, enthusiasm among IT leaders hasn’t translated into full-fledged confidence in their organizations’ ability to execute AI plans. Around a third of respondents (36%) flagged their organizations’ AI maturity as “nascent” or “emerging,” indicating they **may be playing catch-up** when it comes to building out their AI capabilities [Fig. 4].

In addition, while about half of respondents (53%) are extremely confident in their organizations’ ability to execute AI roadmaps, close to half (46%) do express some level of doubt [Fig. 5]. The percentage of doubters is **higher among smaller organizations**: 49% of respondents from companies with $101-$500 million in revenue and 51% of respondents from companies with $501 million-$2 billion in revenue said they were only somewhat to not at all confident, compared to only 40% of respondents from companies with over $2 billion in revenue.

FIG. 3

**Which of the following best describes your current attitude toward implementation of AI applications and initiatives in your organization?** Select no more than three.

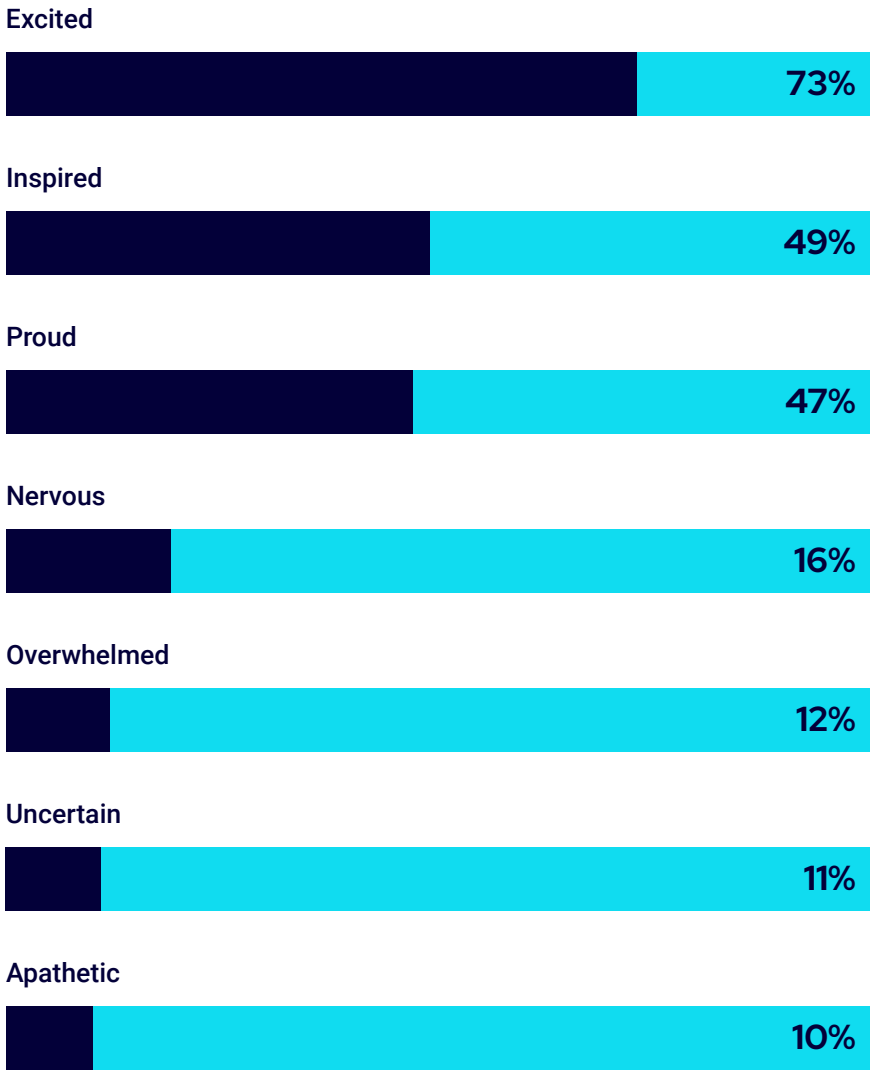
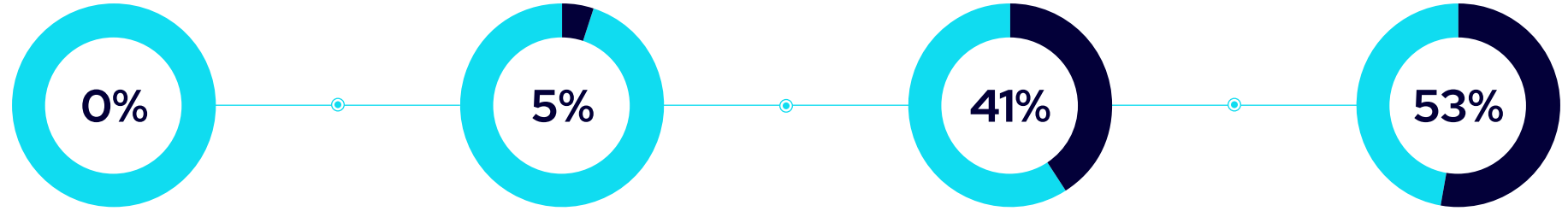


FIG. 5

## Which of the following best describes your confidence level in your organization's ability to execute its AI roadmap?




I'm not at all confident

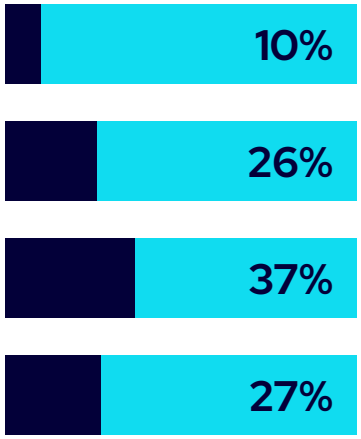
I'm not very confident

I'm somewhat confident

I'm extremely confident

FIG. 4

## Which of the following best describes the current state of AI at your organization?




**Nascent** — We are just beginning to integrate AI into our applications

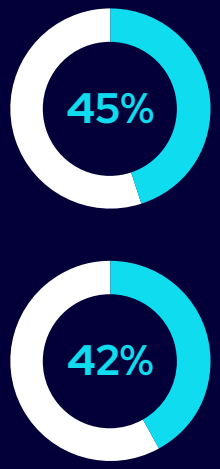
**Emerging** — We are scaling use of AI in our applications

**Mature** — AI is natively integrated into our applications where it makes sense

**Leading-edge** — We're innovating new ways to use AI in our applications

## There will be significant consequences if AI roadmaps aren't achieved

The stakes of AI infrastructure investments are high. If organizations don't achieve the goals laid out in their AI roadmaps, **93% of respondents said there would be consequences** [Fig. 6].




Nearly half of respondents (**45%**) said they will struggle to innovate or have to delay plans for new products or services

**42%** of respondents said they would lose market share to competitors

FIG. 6

## What are the consequences if your organization does not achieve the goals laid out in its AI roadmap? Select all that apply.

We will struggle to innovate — 45%

We will have to put plans for new products or services on hold — 45%

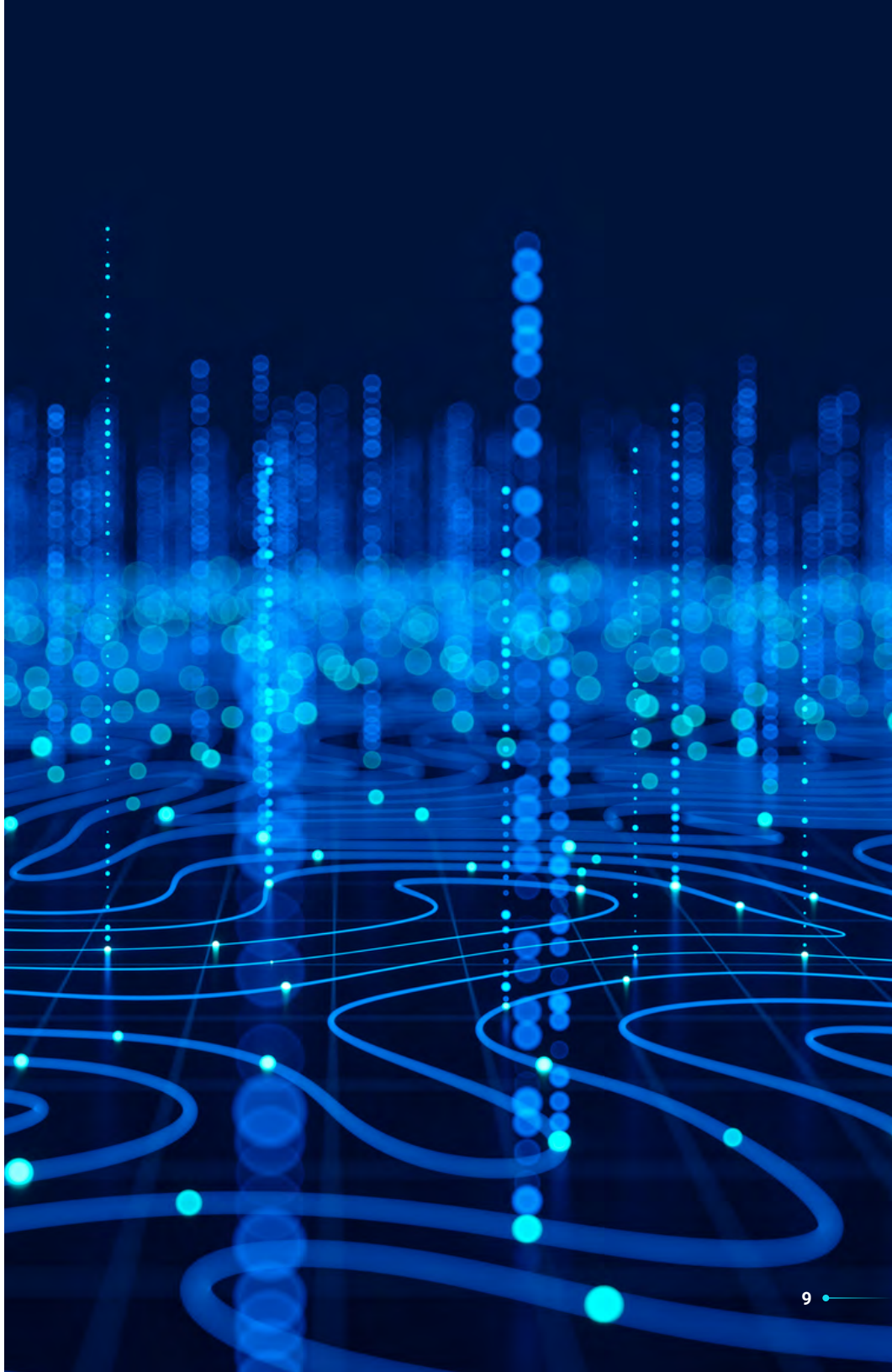
We will be unable to meet employees' demands for digital experience — 42%

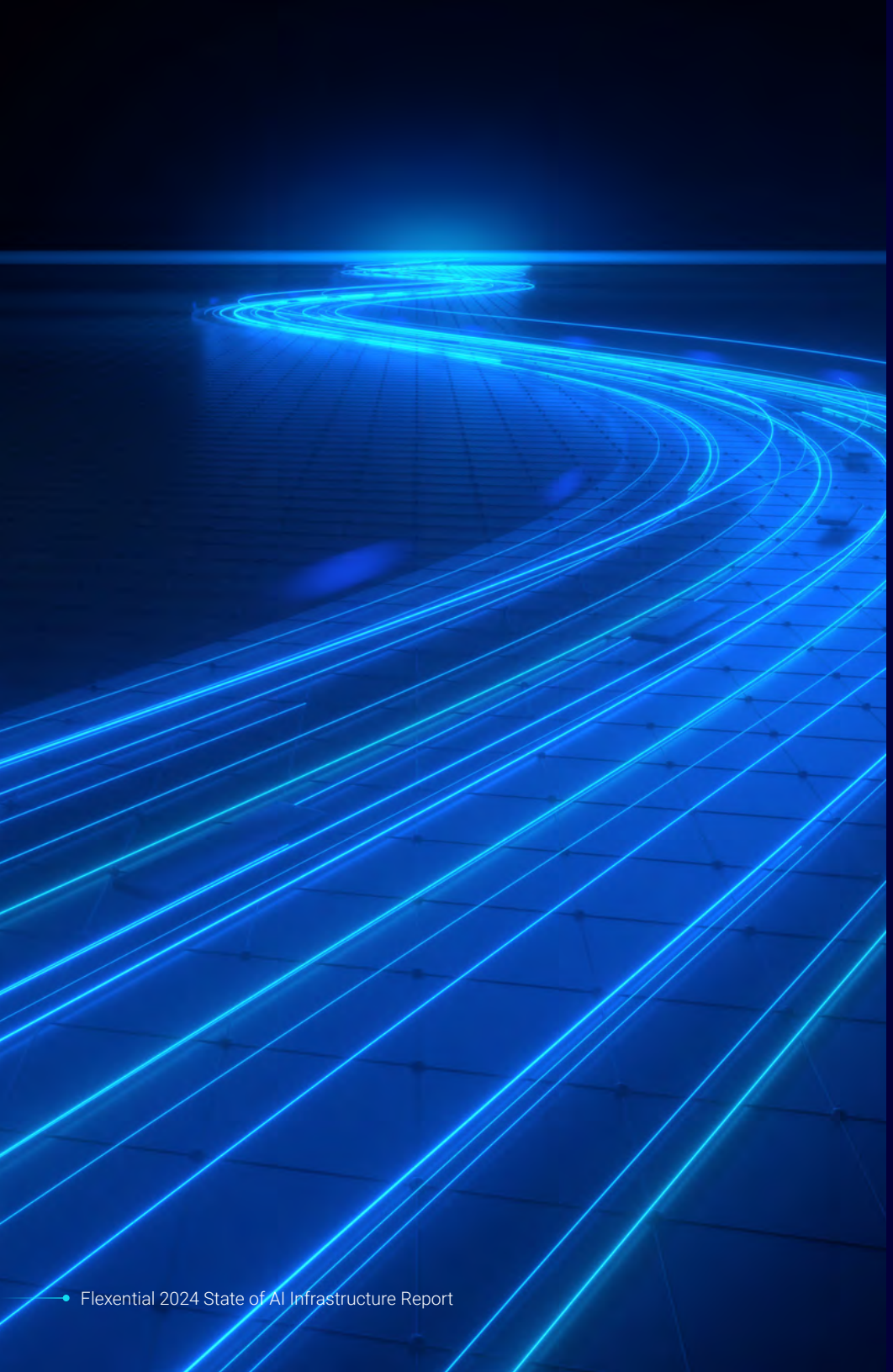
We will lose market share to competitors — 42%

We will be unable to meet customers' demands for digital experience — 41%

We will have a negative impact on the environment — 17%

N/A — there won't be any consequences if my organization does not achieve the goals laid out in our AI roadmap — 7%





# Key takeaways

01

IT leaders have a **board-level mandate** to invest significant resources in executing their AI roadmaps. But while they're enthusiastic about these efforts, IT leaders are less confident about achieving them.

02

Pressure to minimize time-to-revenue clashes with the reality of the **significant infrastructure investments** required to support complex AI use cases moving forward — a key element of most organizations' AI roadmaps.

# Skills gaps, performance issues and data privacy and security concerns pose significant barriers to organizations achieving their AI goals

## Major skills gaps hamper AI progress, especially around high-density computing infrastructure

Building out AI infrastructure is the top priority for organizations' AI roadmaps (see Section 1). However, most organizations **struggle to hire staff** with the skills to support it.

Almost all respondents (91%) report their organization has experienced some sort of skills or staffing gap related to AI in the past 12 months [Fig. 7]. **In particular, more than half of respondents (53%) reported skills gaps or staffing shortages related to the management of specialized computing infrastructure.**

With these staffing challenges in mind, organizations will likely need to utilize AI-enabled private cloud services, specialized technical consultants, or other **outside resources** to achieve ambitious goals in a timely manner.

FIG. 7

**In the past 12 months, has your organization encountered skills or staffing gaps in any of the following areas related to AI?** Select all that apply.

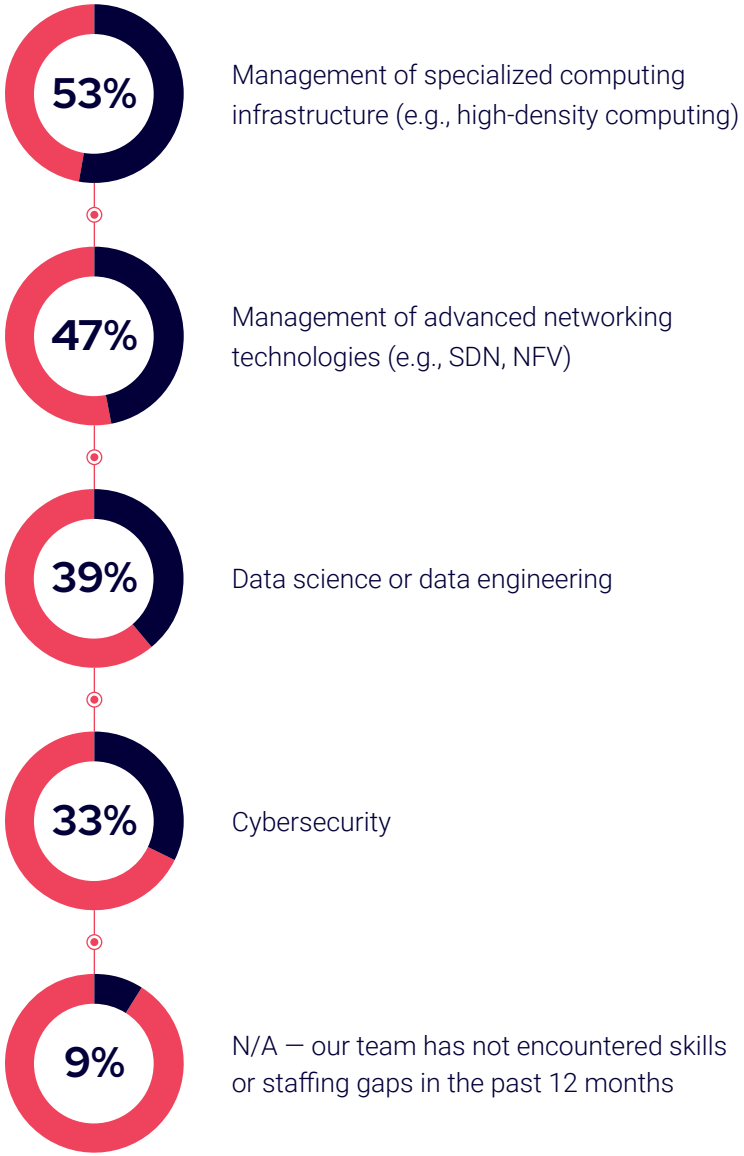
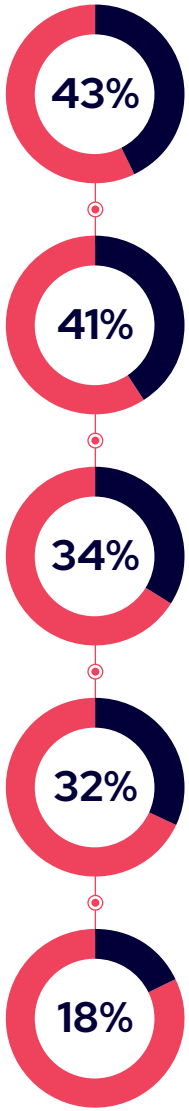


FIG. 8

## In the past 12 months, have you encountered any of the following performance issues with your AI applications or workloads? Select all that apply.




Bandwidth shortages

Unreliable connections

Difficulty scaling data center space and power to meet AI workload requirements

Excessive latency

None of the above — we have not encountered any performance issues with our AI applications or workloads in the past 12 months

## Networking challenges and data center scale are leading causes of AI performance issues

Over three-quarters of respondents (82%) have encountered **some kind of performance issue** with their AI workloads in the past 12 months. Bandwidth shortages were the most common issue cited (43%), followed by unreliable connections (41%) and difficulty scaling data center space and power (34%) [Fig. 8]. Hurdles like these reduce the efficiency of AI applications and increase time to revenue — **impeding organizations' progress on their AI roadmaps, resulting in struggles to innovate, delays launching new products, and other negative consequences discussed in Section 1.**

Adding to the complexity, **C-suite executives may not be fully aware** of problems happening on the ground. For example, 33% of C-suite respondents said their organizations had encountered no AI performance issues over the past 12 months, compared to only 19% of directors and 8% of VPs. Almost a quarter of C-suite respondents (22%) said their organizations had encountered no skills or staffing gaps over the past 12 months, compared to only 7% of directors and 6% of VPs.

## Organizations are pulling back AI applications and workloads from public cloud

A majority of respondents (60%) have had to pull back AI applications or workloads from public cloud to private cloud, on-premise data center or third-party data center in the past 12 months. While sustainability and data privacy/security were top concerns among those who had to pull back AI applications or workloads, 38% said the main consideration was **improving general application performance** [Fig. 9]. Overall, this trend toward moving AI workloads off public cloud will likely exacerbate data center scaling challenges, especially for organizations transitioning to local deployments.

Greater C-suite involvement in driving AI initiatives at large organizations (as flagged in Section 1) may also be driving an **increased focus on infrastructure costs**. Among respondents at organizations with more than $2 billion in revenue that pulled back AI applications or workloads, nearly half (48%) say increasing cost efficiency was a top reason, compared to less than a third of respondents at smaller organizations [Fig. 9a].

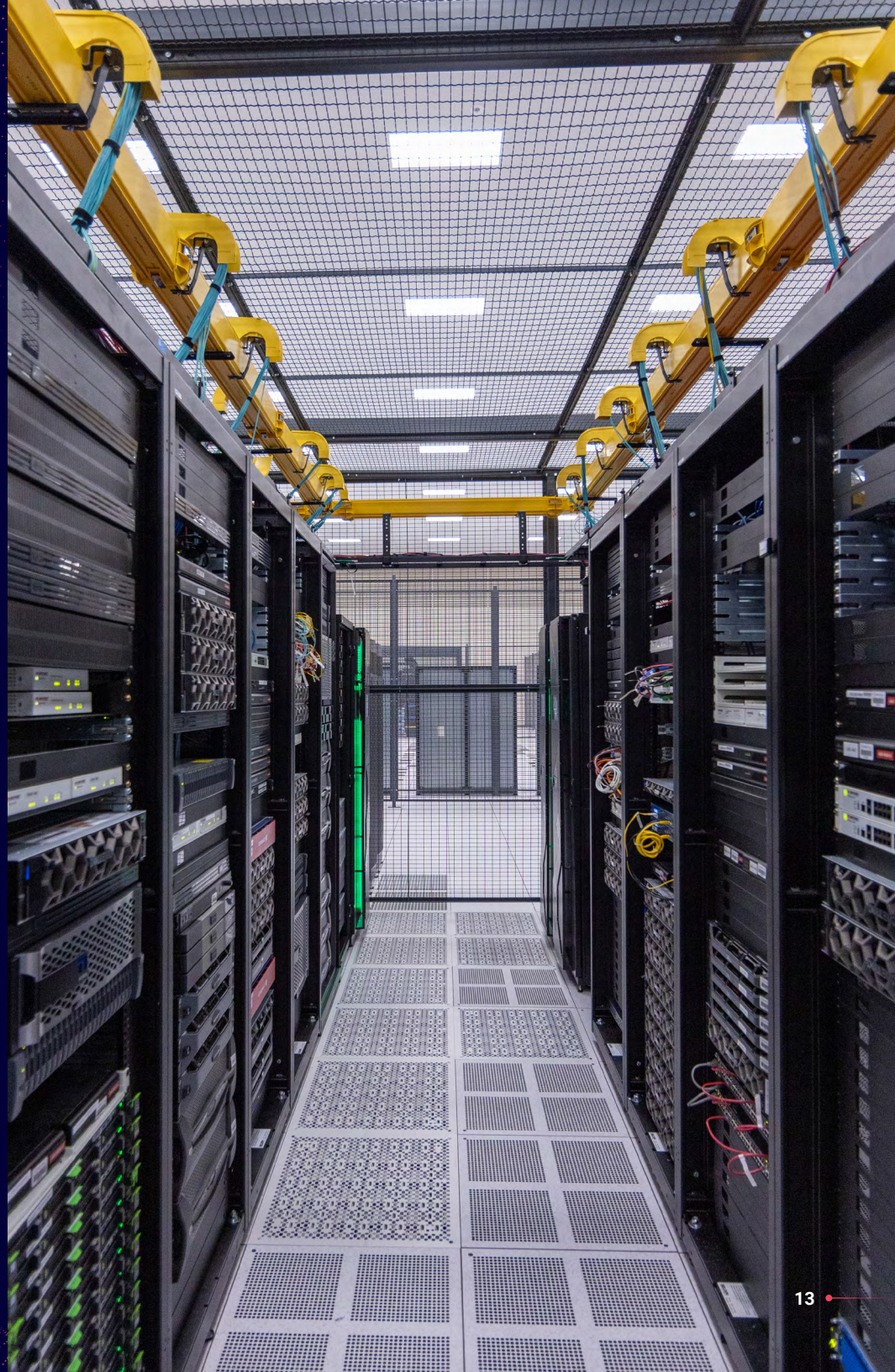


FIG. 9

## What were the primary reasons your organization pulled back AI applications or workloads from public cloud to private cloud and/or an on-premises data center, or third-party data center? Select no more than three.

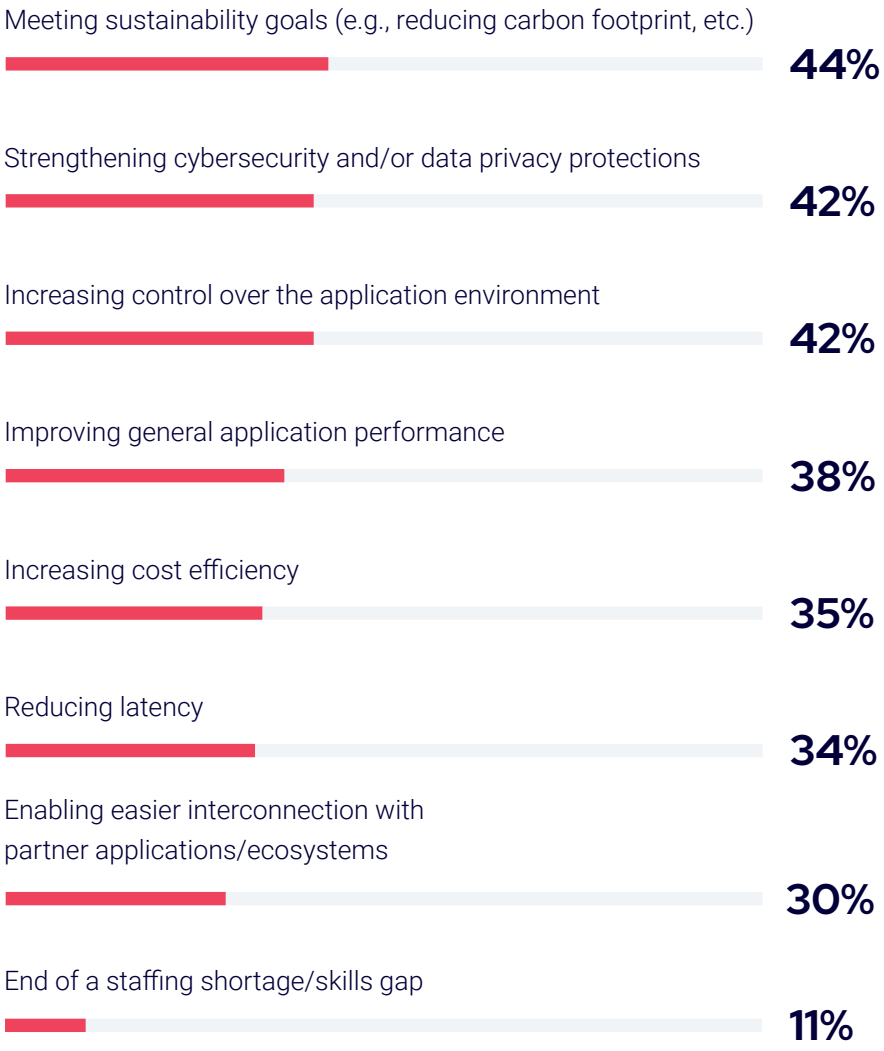



FIG. 9A

## What were the primary reasons your organization pulled back AI applications or workloads from public cloud to private cloud and/or an on-premises data center, or third-party data center? Select no more than three.

| Company revenue | Percentage of respondents selecting "increasing cost efficiency" |
|---|---|
| $101M - $500M | 31% |
| $501M - $2B | 30% |
| > $2B | 48% |

## Data privacy and security concerns related to AI are top of mind

Nearly all respondents (95%) believe their organizations' increased investment in AI has also **increased its vulnerability to cyberthreats** in some way. In particular, around half of respondents (51%) believe storing sensitive data in a different place has increased their organizations' cybersecurity vulnerability. For organizations using or building on public large language models (LLMs) like ChatGPT, **upstream data privacy concerns** are likely a major factor, as these models often lack clear privacy and security policies.

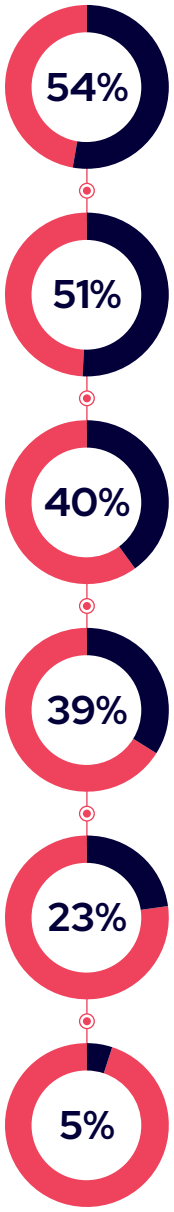
Most shockingly, 40% of respondents said their cybersecurity teams **don't understand how to protect AI applications and workloads** [Fig. 10]. Beyond technical skills gaps, this likely also reflects a lack of knowledge around how to build or implement platforms compliant with major data privacy legislation such as the EU's General Data Privacy Regulation (GDPR). **That's a major skills gap organizations will need to fill with third-party expertise and infrastructure, especially as investments in AI continue to grow.**

How are organizations currently addressing these concerns? Strengthening data privacy and security protections was one of the top reasons organizations pulled back AI applications and workloads from the public cloud (42%) [Fig. 9]. Many likely made this move to **keep sensitive data under tighter control** within an on-premise or third-party data center/private cloud.

FIG. 10

### How has increasing your organization's investment in AI increased its vulnerability to cyberthreats?

Select all that apply.




The complexity of AI applications creates a larger attack surface which is inherently more vulnerable

We're storing sensitive data in a different place than before (e.g., closer to the edge of the network)

Our cybersecurity team doesn't understand how to protect AI applications and workloads

We're storing and processing more sensitive data than before

There are more employees involved with managing our AI applications, leading to more weak spots

N/A — my organization's investment in AI has not increased its vulnerability to cyberthreats


# Key takeaways

01

Multiple challenges stand in the way of organizations' ability to execute their AI roadmaps. **Skills gaps** make it difficult to meet a pressing need for high-density infrastructure, while **performance issues** tied to networking and data center scaling difficulties hold back AI initiatives' time to revenue. At the same time, IT leaders **lack trust in their own cybersecurity teams' expertise** to answer AI's unique data privacy and security challenges, and keeping up with data demand and processing is a concern.

02

Most organizations will need to draw on **third-party expertise and specialized infrastructure** to fill these gaps and meet ambitious AI goals. Adopting flexible solutions is key as needs continue to change.

SECTION 3

# Organizations are leveraging colocation to meet AI infrastructure challenges, but there's room for additional optimization

## When deciding where to deploy AI workloads, security, cost efficiency and application performance are top considerations

With hybrid cloud infrastructure, strategically choosing where to deploy **compute-heavy AI workloads** is important for maximizing performance, efficiency and other goals. When respondents make these key decisions, cybersecurity and data privacy are their top consideration (62%), followed by cost efficiency (44%) and general application performance (40%) [Fig. 11]. These factors mirror the reasons why organizations have **pulled back workloads from the public cloud** (see Section 2), signaling a consistent approach across use cases.

When it comes to latency, larger organizations are more sensitive than smaller ones. Only about a quarter of respondents at organizations with $101-$500 million in revenue (27%) say **low latency is a top consideration** when deciding where to deploy AI workloads, compared to 41% of respondents at organizations with $501 million-$2 billion in revenue and 35% of organizations with $2 billion in revenue or more [Fig 11a].

FIG. 11

## When considering where to deploy an AI workload on the cloud (i.e., private versus public cloud), which of the following factors are most important?

Select no more than three.

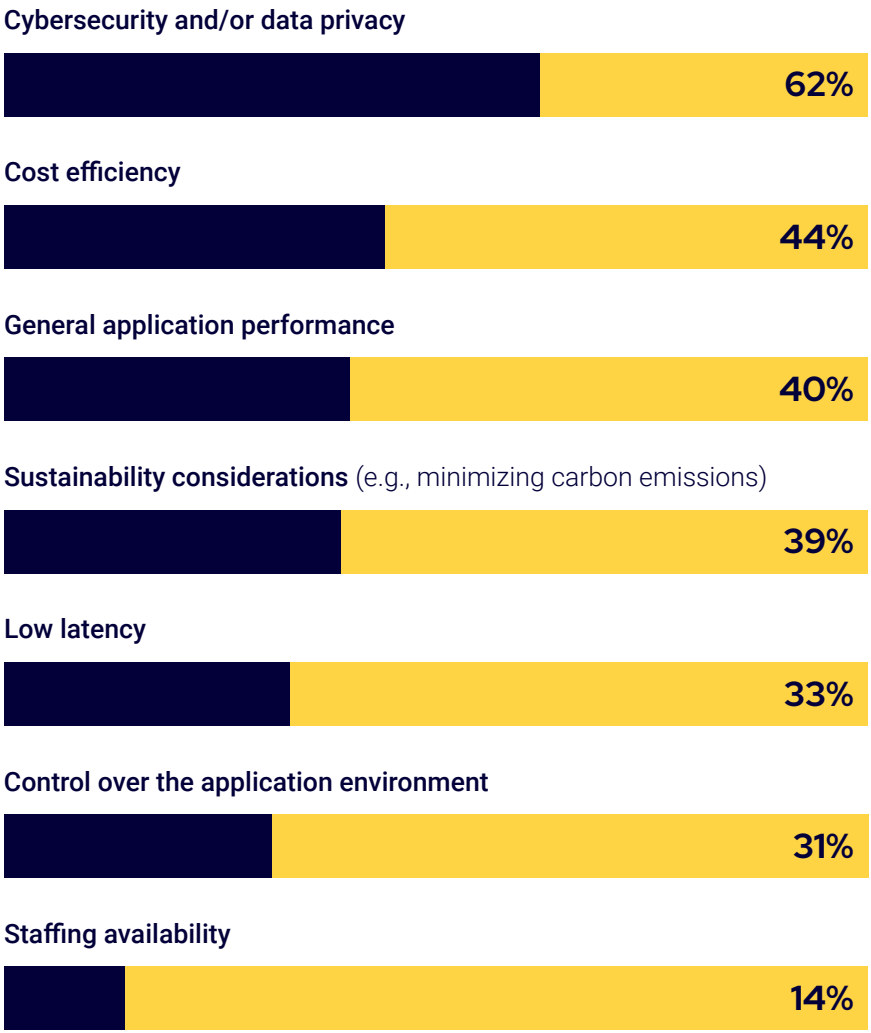



FIG. 11A

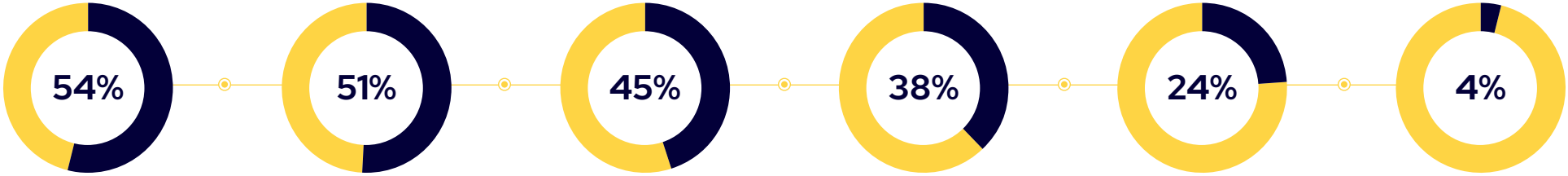
## When considering where to deploy an AI workload on the cloud (i.e., private versus public cloud), which of the following factors are most important?

Select no more than three.

| Company revenue | Percentage of respondents selecting "low latency" |
|---|---|
| $101M - $500M | 27% |
| $501M - $2B | 41% |
| > $2B | 35% |

FIG. 12

## Which of the following tactics is your organization currently implementing to reduce performance issues for its AI applications or workloads? Select all that apply.




Using 5G networks (e.g., 5G-enabled IoT devices, etc.)

Using third party colocation data centers to process data closer to the edge of the network

Using network function virtualization (NFV)

Implementing software-defined networking (SDN)

Deploying AI hardware on-premise

None of the above

FIG. 13

## Where is the data you utilize or plan to utilize for AI training or inference applications currently housed? Select all that apply.


## Organizations rely heavily on third-party colocation data centers to fulfill AI-related infrastructure needs

AI applications and workloads demand **low latency and specialized infrastructure** to function optimally. Meeting these needs on-premise is difficult. Unsurprisingly, less than a quarter of respondents (24%) say they're deploying AI hardware on-premise [Fig. 12]. Instead, over half of respondents (51%) say they're **leasing rack space in third-party colocation data centers to process data closer to the edge of the network.** When asked about the data they utilize or plan to utilize for AI training or inference application, most respondents also said this data was housed in their colocation data center (60%) [Fig. 13].

Shifting compute and data storage to the edge of the network inherently minimizes latency. In addition, many colocation data centers offer **even more efficiencies** — e.g., by enabling the deployment of GPU hardware next to CPU stacks and key data lakes — that further improve performance. Efficiency in turn reduces costs, which are always a major concern for AI workloads. In addition, leveraging colocation **avoids public cloud network fees**, improving time to value and boosting results.

## However, organizations could use colocation and other AI infrastructure optimization strategies more effectively

While most respondents say they're storing and processing data in third-party colocation data centers closer to the edge of the network, less than a quarter (24%) said they deployed the most GPUs there [Fig. 14]. This suggests organizations are **not yet deploying their heaviest AI workloads at the edge**, which could limit their performance benefits.

As businesses increasingly look to collaborate in real time across disparate IT ecosystems, **interconnection-led infrastructure** is becoming vital. However, our research reveals many organizations' connections are not optimized. Around a third of respondents (34%) say their AI applications access data via the public internet rather than a private connection [Fig. 15]. **Public connections increase latency** and are more vulnerable to cyberattacks and breaches. Their higher potential for network disruption reduces time to value for AI applications.

Shifting to private connections would mitigate these risks. In addition, many AI workloads will utilize a hybrid or multi-cloud approach. It's important that IT infrastructure solutions enable **reliable connectivity across environments** by offering the flexibility to easily integrate multiple network options, including private connections.

FIG. 14

### Which of the following best describes where your organization deploys the most GPUs?

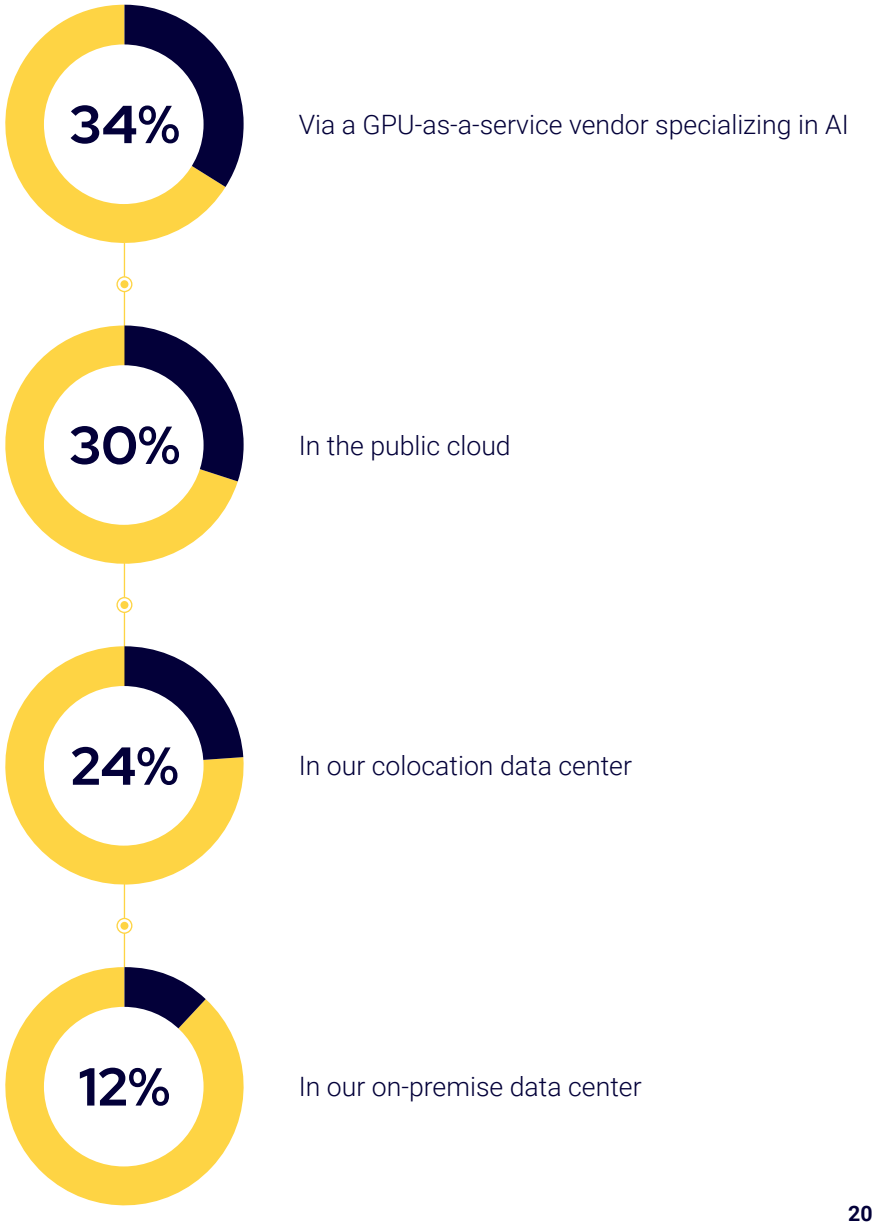
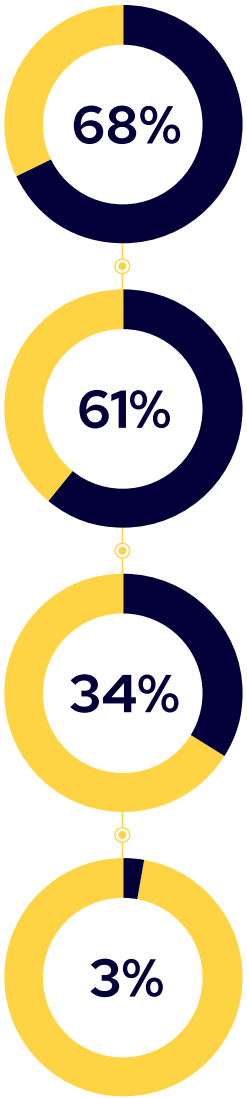


FIG. 15

## How do your AI-driven applications access the data you utilize or plan to utilize for AI training or inference applications? Select all that apply.

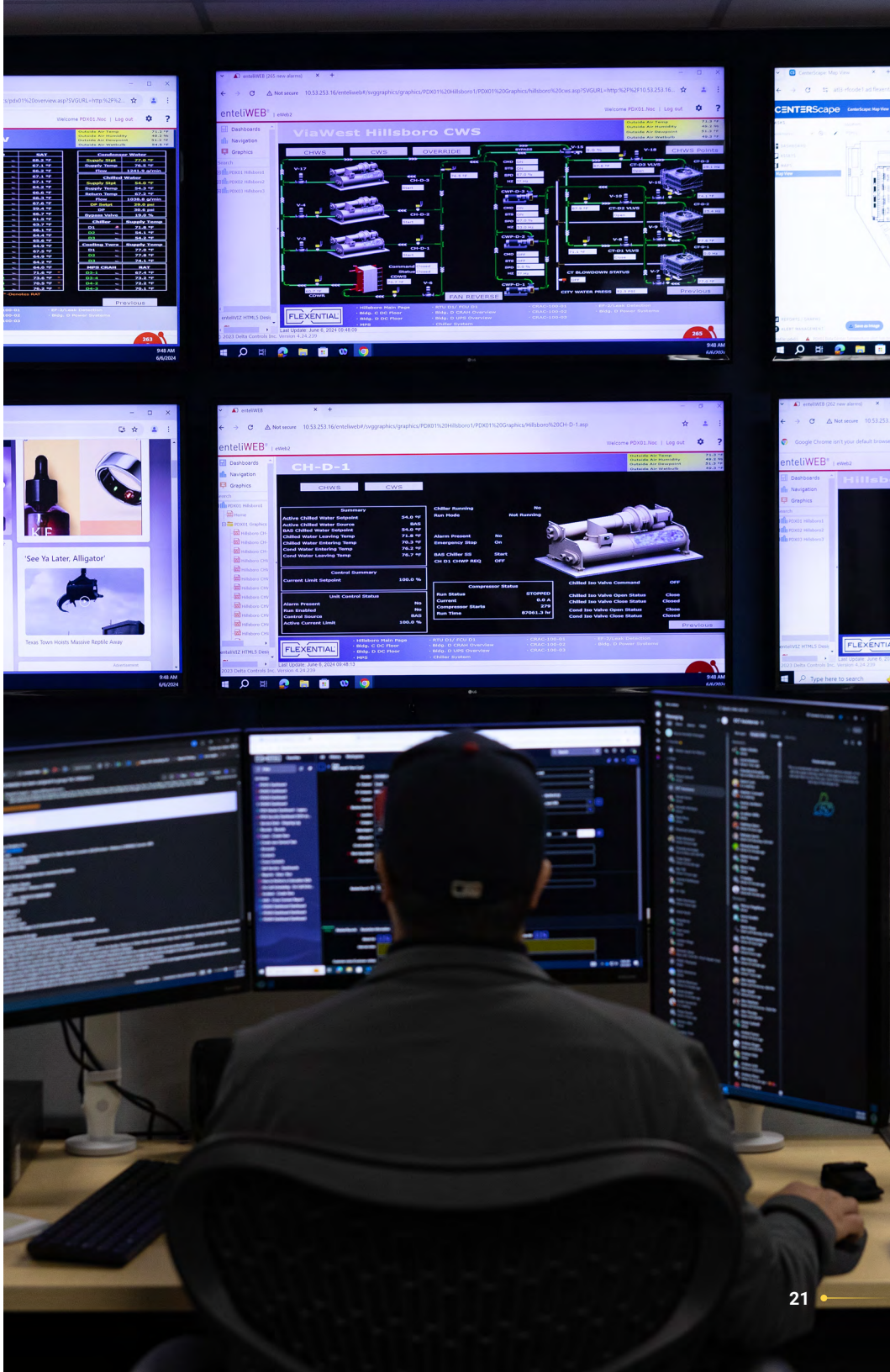



Via private connections within the public cloud

Via private connectivity between colocation data centers or our facilities

Over the public internet

We don't have a way to access all of this data yet



# Key takeaways

01

Strategic decision-making about where to deploy AI workloads is vital for optimizing cost efficiency, security and other concerns. Many organizations already **leverage third-party services** to manage AI infrastructure needs, including colocation data centers.

02

However, organizations can realize even greater value from these investments. To start, **turning public connections into private ones** and moving GPUs and other AI infrastructure into colocation data centers closer to the edge of the network would improve security and performance.

SECTION 4

# Improving data center sustainability is a top priority for IT leaders

## Like AI, IT sustainability has become a board-level issue — and IT leaders are feeling the heat

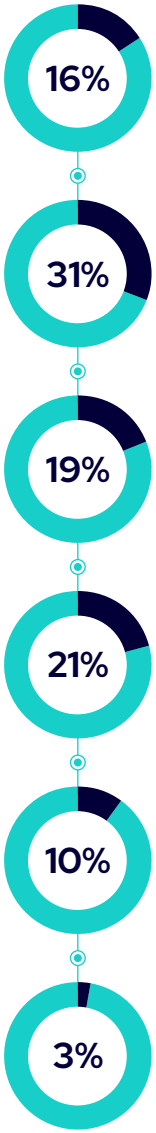
Initiatives to reduce data centers' carbon footprints are not new, **but the AI boom has made the discussion more urgent and visible**. One recent analysis projected that by 2027, AI infrastructure would consume up to 134 terawatt-hours of power annually — roughly the same amount of energy the entire nation of Sweden uses in a year.

Unsurprisingly, almost all respondents (97%) felt some level of pressure to improve IT sustainability [Fig. 16]. Around half (47%) reported feeling **more pressure than they did five years ago.** Respondents at larger organizations with $501 million or more in annual revenue were more likely to say they felt increased pressure to make IT infrastructure more sustainable [Fig. 16a].

Among those who felt pressure, around half said the C-suite (51%) or the board (46%) was one of the driving forces behind that pressure [Fig. 17]. These groups outranked customers (36%), internal sustainability departments (36%) and the general public (29%) as sources of pressure on IT sustainability. **C-suite scrutiny** may be one reason why sustainability was a top reason for pulling back AI workloads from the public cloud (see Section 2).

FIG. 16

## Compared to five years ago, how much pressure do you feel to make IT infrastructure more sustainable (e.g., by reducing its carbon footprint, etc.)?




Significantly more pressure than five years ago

Somewhat more pressure than five years ago

The same amount of pressure as five years ago

Somewhat less pressure than five years ago

Significantly less pressure than five years ago

N/A — I don't feel any pressure to make IT infrastructure more sustainable

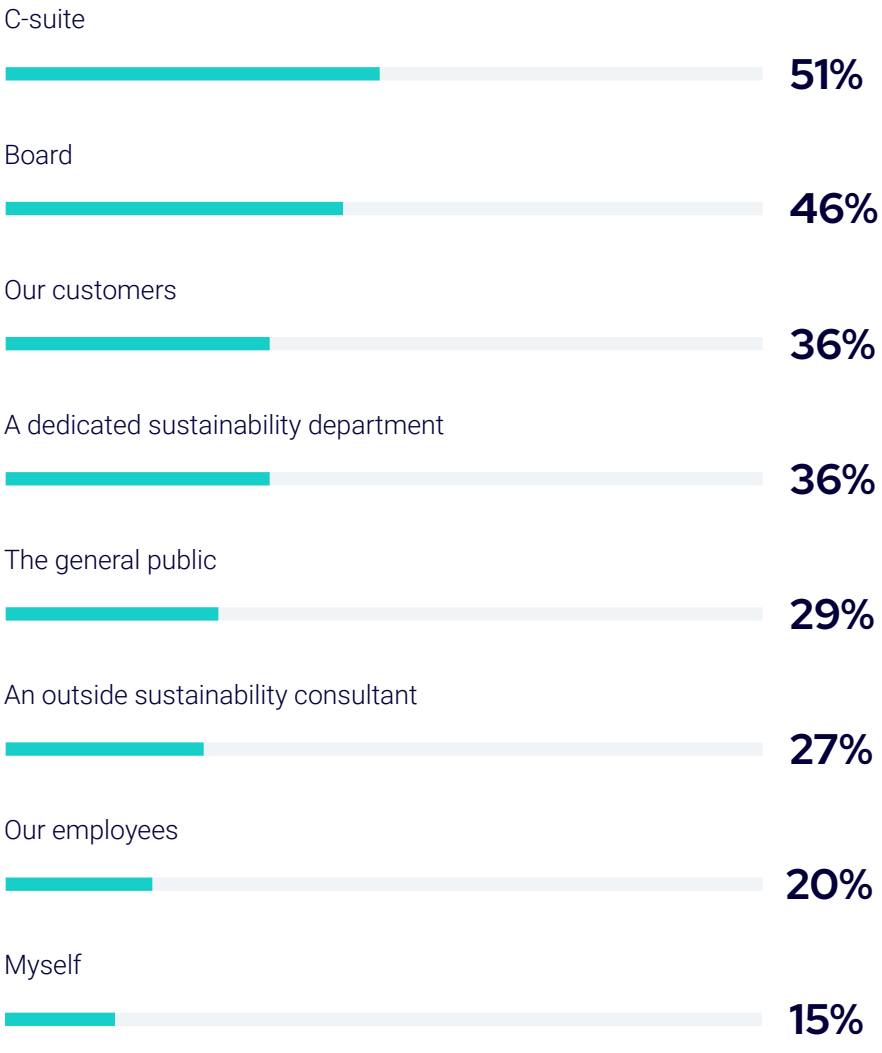
FIG. 16A

### Compared to five years ago, how much pressure do you feel to make IT infrastructure more sustainable (e.g., by reducing its carbon footprint, etc.)?

| Company revenue | Percentage of respondents selecting "significantly more" or "somewhat more pressure" |
|---|---|
| $101M - $500M | 43% |
| $501M - $2B | 50% |
| > $2B | 51% |

FIG. 17

## Which of the following groups or individuals are the driving force behind the pressure you feel to make your organization's IT infrastructure more sustainable? Select no more than three.



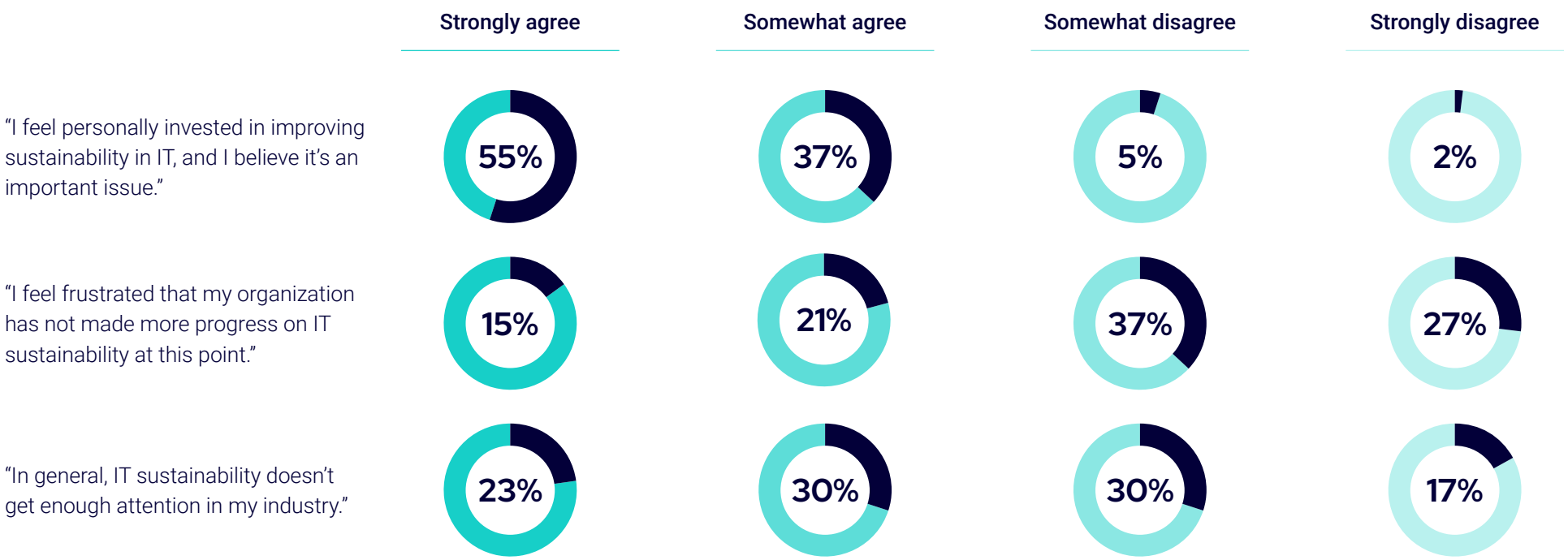
## IT leaders are personally invested in improving the sustainability of their IT infrastructure

Nearly all respondents (93%) feel at least somewhat personally invested in improving sustainability in IT, and a majority (53%) believe the issue doesn't get enough attention [Fig. 18]. However, most respondents (63%) are not frustrated by their organizations' progress with IT sustainability, indicating that they're **satisfied with their organizations' current efforts.**

In line with other respondents' perception of the sources of pressure around IT sustainability, **C-suite respondents were particularly passionate** about the issue. Over two-thirds (70%) strongly agreed that they feel personally invested in improving sustainability in IT, and 41% strongly agreed that IT sustainability doesn't get enough attention in their industry.

FIG. 18

### How much do you agree or disagree with each of the following statements about sustainability initiatives in IT?




[1]Note: due to rounding, not all percentages in this chart add up to 100%.

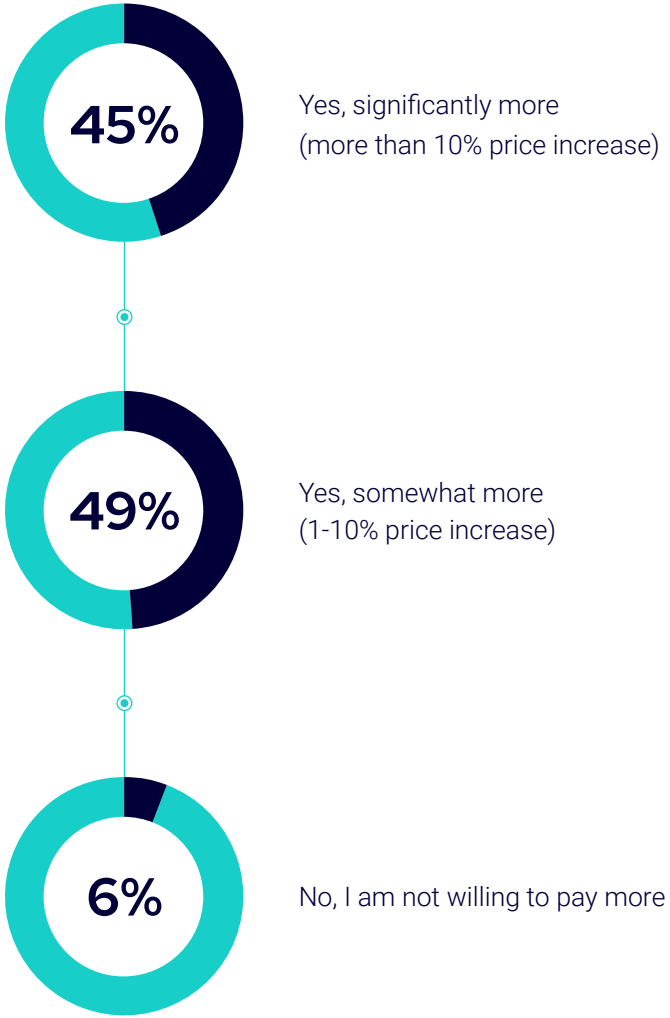
## When it comes to data center sustainability, IT leaders are willing to pay a premium

Third-party data centers or cloud vendors can often deploy sophisticated technology to reduce their carbon footprints, such as fungible air delivery systems that can customize cooling to the individual room, row or even rack level. Nearly all respondents (94%) **would pay a premium for more sustainable cloud services**, and almost half (45%) would pay more than a 10% increase [Fig. 19].

This trend was particularly pronounced among **retail industry respondents**, over half of whom (52%) said they would be willing to pay more than a 10% increase. It's possible that retail IT leaders are more willing to pay because they're more likely to reap a return on their investment by marketing their companies as "green" to consumers.

FIG. 19

**Are you willing to pay higher costs for your data centers or third-party cloud vendors to use clean or renewable energy and/or buy credits to offset their carbon footprints?**



# Key takeaways

01

Increasing investment in **energy-intensive AI infrastructure** means increasing focus on another C-suite priority: **improving data center sustainability.**

02

To bridge this gap, IT leaders are willing to **pay a significant premium** for better sustainability outcomes from third-party data centers or cloud vendors. This move isn't inspired by C-suite and board pressure alone: **IT leaders themselves** are invested in reducing data centers' carbon footprints, too.

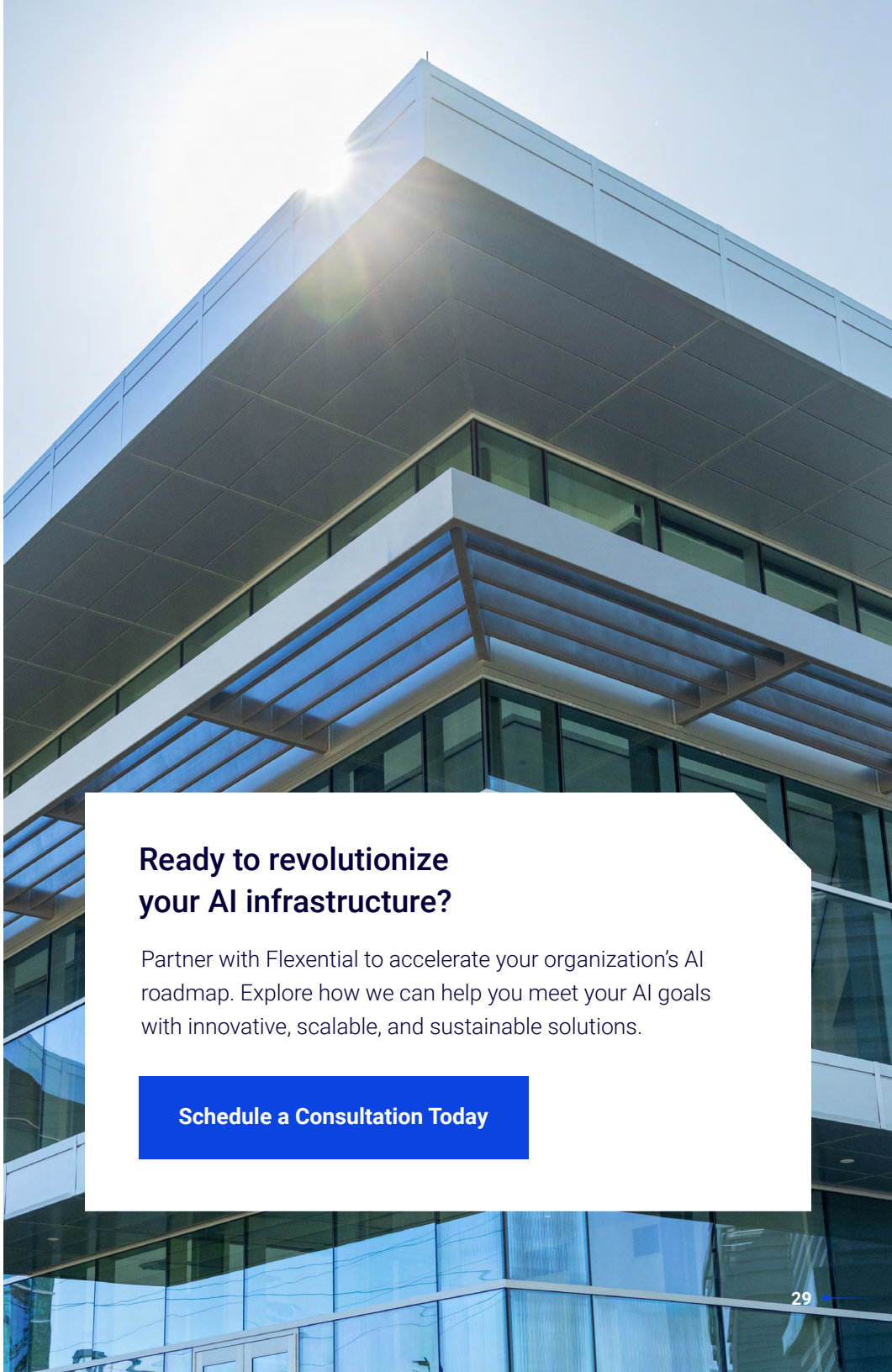
CONCLUSION

# It’s time for a fresh approach to AI infrastructure

As AI accelerates the pace and scope of organizational change and innovation, IT leaders must **reimagine how they architect infrastructure** to achieve ambitious business goals. That means:

- Sourcing **high-density compute capacity** that can scale with AI workloads, while flexibly integrating new tools and technologies as their organizations’ needs evolve.
- Securing the sensitive data processed in AI applications through **robust security measures** that extend throughout their entire IT architecture, ensuring compliance with ever-evolving privacy rules.
- Leveraging **software-led interconnection** to support real-time collaboration across IT ecosystems and large, distributed workloads.
- Deploying **liquid-cooling technologies** and other specialized equipment and strategies to address C-level sustainability priorities without sacrificing performance.

Promoting innovation, rolling out new products and services, maintaining or increasing market share — all of these priorities depend on organizations’ abilities to make these core infrastructure investments. An **experienced data center partner** can swiftly interconnect and optimize your IT infrastructure to deliver the high bandwidth, consistent throughput and low-latency private connections you need to scale your most crucial AI workloads.



## Ready to revolutionize your AI infrastructure?

Partner with Flexential to accelerate your organization’s AI roadmap. Explore how we can help you meet your AI goals with innovative, scalable, and sustainable solutions.

**Schedule a Consultation Today**

# Methodology

In March and April 2024, Flexential surveyed 350 IT decision-makers at the director level or above at organizations with over $100 million in annual revenue. All respondents had knowledge of their organizations' AI implementation and related infrastructure build-outs. Respondents came from a range of industries.

| Annual revenue | |
|---|---|
| $101M - $500M | **43%** |
| $501M - $2B | **29%** |
| > $2B | **29%** |

| Job level | |
|---|---|
| Director | **60%** |
| Vice President | **25%** |
| C-suite | **15%** |

| Industry | |
|---|---|
| Technology/IT Services/Software | **23%** |
| Manufacturing | **14%** |
| Financial Services | **12%** |
| Retail/Wholesale | **9%** |
| Healthcare | **8%** |
| Telecom/Networks | **6%** |
| Education | **5%** |
| Entertainment/Media | **5%** |
| Business and Professional Services/B2B Services | **4%** |
| Transportation/Logistics | **4%** |
| Construction | **4%** |
| Nonprofit | **2%** |
| Legal | **2%** |
| Utilities | **1%** |